Using principal curves to analyze traffic patterns on freeways

Jochen Einbeck* and Jo Dwyer
Durham University, Department of Mathematical Sciences,
Science Laboratories, South Road,
Durham, UK

Abstract

Scatterplots of traffic speed versus flow have caught considerable attention over the last decades due to their characteristic half-moon like shape. Modelling data of this type is difficult as both variables are actually not a function of each other in the sense of causality, but are rather jointly generated by a third latent variable, which is a monotone function of the traffic density. We propose local principal curves as a tool to describe and model speed-flow data, which takes this viewpoint into account. We introduce the concept of calibration curves to determine the relationship between the latent variable (represented by the parametrization of the principal curve) and the traffic density. We apply local principal curves to a variety of speed-flow diagrams from Californian freeways, including some so far unreported patterns.

Key Words: fundamental diagram; capacity; local principal curves; smoothing.

*jochen.einbeck@durham.ac.uk

1 Introduction

Scatterplots of speed versus flow have been widely analyzed and discussed in transportation science, and have recently attracted new interest with the rapid advances in the development of Intelligent Transportation Systems. As an example, consider data plotted in figure 1 (left), recorded on 10th July 2007 (00:00 to 23:59) on the Californian Freeway I280-N, Lane 1, VDS ("vehicle detector station") number 716450. The data show a characteristic and frequently reported half-moon like shape. Roughly, the upper and the lower cluster correspond to uncongested and congested operating condition, respectively, and the few data points between them to an unstable transition region. Based on a cluster analysis, Xia & Chen (2007) argued that actually five different operating conditions should be distinguished.

Under equilibrium conditions, i.e. stationary speed and spatially homogeneous density, it is well known that the speed v and the flow q are related through the fundamental identity $q = kv$, where k is the traffic density. The association between speed, flow and density is often referred to as the fundamental diagram. As Wu (2002) points out, the fundamental identity specifies the fundamental diagram only up to one degree of freedom. In other words, one has to impose an additional constraint on any pair of the three variables in order to specify the fundamental diagram fully. This is usually achieved by fixing the $k - v$ relationship. For instance, the original model suggested by Greenshields (1935) uses $k(v) = k_j(1 - v/v_f)$, where k_j is the jam density corresponding to $v = 0$ and and v_f the free-flow speed. Having fixed the speed-density relationship, the speed-flow relationship is determined by $q(v) = k(v)v$, which in the special case of the Greenshields model (hereafter: GM) takes the shape $q(v) = k_j v(1 - v/v_f)$, i.e. a parabola without an intercept. Several other, generally more complex, functional relationships between k and v have been proposed since

then, see e.g. Kockelman (2001) or Wu (2002) for an overview on this literature. An interesting and early reference comparing different speed-density models from a statistical point of view is Drake, Schoefer & May (1967). More recently, Wu (2002) proposed to avoid the usually applied "trial and error" model selection strategy by relating the parameters of the fundamental diagram to microscopic road parameters. The reason why the $k - v$ relationship is preferred to any other pair of variables is simply that this is the only one which is monotonic. This is illustrated in figure 1 (right), using here occupancy, the quantity returned by default by PeMS, which is roughly linearly related to density (Hall, 2002).

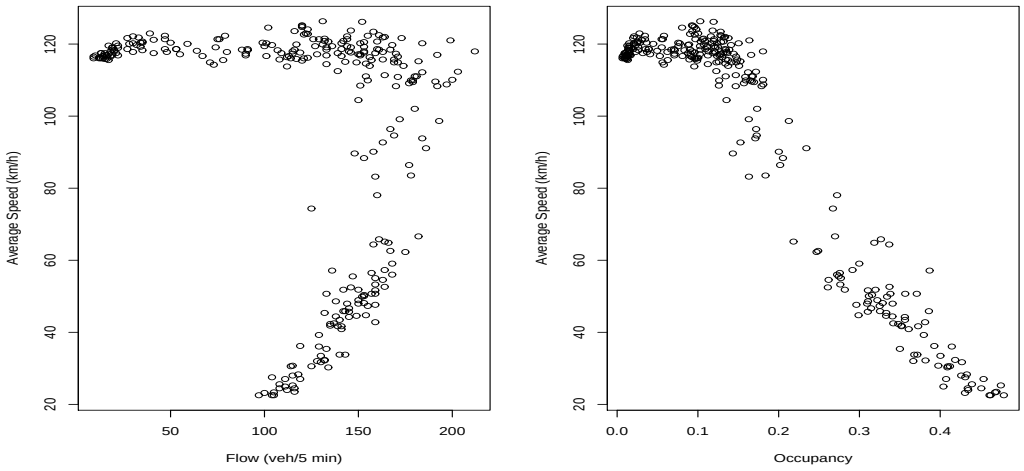


Figure 1: Fundamental diagram recorded on Freeway I280-N; left: speed-flow; right: speed-occupancy.

There have hardly been made any attempts at modelling the $q - v$ relationship directly (rather than through the $k - v$ relationship); Li (2008) mentions one instance used in the Highway capacity manual 2000. One reason for this reluctance may be that any functional form between q and v is hard to justify. Obviously, v cannot be

seen as a function of q as we have potentially two different outputs for the same input. But also the other way round, $q = q(v)$, seems somewhat contrived: speed v is quite difficult to measure while the flow q is very easy to measure — realistically, nobody would be interested in predicting flow from speed. Also, traffic flow is not a function of speed in the sense of causality, it is rather that drivers have to obey the constraints set by the current road conditions, and this will affect both speed and flow. As a consequence, it seems more natural to consider both variables as the two-dimensional output of a function $\begin{pmatrix} q \\ v \end{pmatrix}(t)$ of some (latent) variable, say t. This also does the job of fixing the remaining degree of freedom in the fundamental diagram, but it implies a symmetric view on the variables; the resulting model is invariant w.r.t. interchanging the coordinate axes for q and v.

The statistical concept corresponding to this viewpoint is a *principal curve*: a smooth curve passing through the "middle of the data cloud". Principal curves were introduced by Hastie & Stuetzle (1989) (hereafter: HS) as a nonparametric extension to linear principal component analysis. Chen, Zhang, Tang & Wang (2004) have applied HS principal curves to speed-flow data and showed that this leads generally to better fits than the Greenshields-type parametric models of flow given speed. We will take things on from here and illustrate the benefits and relevance of principal curves in the context of the fundamental diagram. The methodology that we will be using for the actual curve fitting is that of *local principal curves* (Einbeck, Tutz & Evers, 2005b, hereafter LPC). In Section 2, we explain briefly how LPCs work, and we demonstrate that the curve parametrization, representing the latent variable, is a monotonic function of the traffic density, which was singled out as "the primary factor to define the level of service on a freeway" by Xia & Chen (2007). Specifically, we introduce the novel concept of a calibration curve, which relates the curve parametrization to den-

sity or occupancy. In Section 3, we fit local principal curves to a temporal sequence of speed-flow diagrams, compare them with HS and Greenshields curves, and link them to the notion of capacity. We further give an overview of characteristic speed-flow patterns collected in the course of our studies, some of which do not follow the "classical" shape of the fundamental diagram. We finish with a conclusion in Section 4. The techniques proposed in this paper are still feasible even if no measured values of density or occupancy have been provided. Hence, when we use occasionally the term *speed-flow diagram* instead of fundamental diagram, we emphasize that we do have the speed and flow values, but not necessarily anything else, available.

2 Principal curves and the fundamental diagram

Principal curves are smooth curves passing through the middle of the distribution of a data cloud. Several competing concepts of principal curves have been developed over the last 20 years, which can be divided into two major families: (i) "Top-down" methods, such as HS curves, start with some straight line, which is then iteratively bent until it fits satisfactorily through the data cloud and some global error criterion is minimized; (ii) "Bottom-up" methods construct the principal curve point by point, at each iteration only considering the information provided in the local neighborhood. Bottom up methods feature generally a larger flexibility at the expense of higher variability. A representant of this family are local principal curves (Einbeck, Tutz & Evers, 2005b):

2.1 Local principal curves

To fix terms, let $K_H(\cdot) = |H|^{-1/2} K\left(H^{-1/2}\,\cdot\right)$, with $K(\cdot)$ being a 2−dimensional kernel function (e.g. a two-dimensional Gaussian density), $H = \mathrm{diag}(h_1^2, h_2^2)$ a multivariate

(2×2) bandwidth matrix, $w_i^x = K_H(x_i - x)/\sum_{i=1}^n K_H(x_i - x)$, and $x_i^T = (x_{i1}, x_{i2})$, $i = 1, \ldots, n$ (think of $x_{i1} = q_i$ as flow-values and $x_{i2} = v_i$ as speed-values). The "middle" of the data cloud is estimated as follows:

Algorithm: Local principal curves (LPC)

1. Select a starting point $x_0 \in \mathbb{R}^2$ and a step size $t_0 > 0$. Set $x = x_0$.
2. Calculate the local center of mass $\mu^x = \sum_{i=1}^n w_i^x \, x_i$ at x. Denote by μ_j^x the j-th component of μ^x, $j = 1, 2$.
3. Estimate the local covariance matrix $\Sigma^x = (\sigma_{jk}^x) \in \mathbb{R}^{2\times 2}$ at x via
$$\sigma_{jk}^x = \sum_{i=1}^n w_i^x (x_{ij} - \mu_j^x)(x_{ik} - \mu_k^x).$$
Let γ^x be the first eigenvector of Σ^x.
4. Setting $x := \mu^x + t_0 \gamma^x$, one finds the updated value of x.
5. Repeat steps 2 to 4 until the sequence of μ^x remains approximately constant (implying that the end of the data cloud is reached). Then set again $x = x_0$, set $\gamma^x := -\gamma^x$ and continue with step 4.

In words, this algorithm does nothing other than compute alternately a local center of mass and a localized first principal component. The series of local centers of mass μ^x make up the local principal curve. For the data from figure 1, they are plotted through "+" symbols in figure 2 (left), and the curve is obtained from them through linear (as done here) or cubic interpolation. This algorithm is implemented in the `lpc` function which runs under R (R Development Core Team, 2009) and is available for download from the source given in the Conclusion of this paper. The step size t_0 and the bandwidths h_1 and h_2 are not chosen automatically as part of this algorithm, but need to be specified. This choice is simplified when both variables operate on the same scales, or when the data are previously scaled e.g. by dividing both variables through their range. In this case it is reasonable to set $h_1 = h_2 \equiv h$,

and to set the step length t_0 equal to h as well. When working with speed-flow data, where speed is measured in km/h and flow in veh/5 min, the range of both variables is of a similar magnitude, so that there is no need for rescaling or for the use of unequal bandwidths. As most fundamental diagrams operate on a similar range of flow and speed values, there is also no need to choose the bandwidth h for each data set anew. The starting point x_0 can be selected at random or by hand, and does not need to be part of the data cloud. The shape of the fitted curve can vary depending on the cluster or branch in which it is situated, but the actual position within the particular branch is of little relevance. In order to obtain the half-moon shape we recommend to place it somewhere in the congested half. Unless otherwise stated, we work with $h = t_0 = 12$ and $x_0 = (150, 50)^T$ for all examples throughout this paper.

Though the LPC algorithm does not optimize a global error criterion, it is right to say that a "good" fit corresponds to one which results in small orthogonal distances between data and curve (unlike vertical ones as in the regression context). Residuals do not play an active role in the construction of the curve but, if desired, they can be retrospectively calculated through the distance between a data point and its orthogonal projection on the curve; see figure 2 (right) for a graphical illustration of this concept.

2.2 Parametrization and Calibration

So far, the local principal curve is simply defined by a set of points. This may be useful for descriptive purposes, but does not yet allow to make use of the curve for further purposes, for instance prediction. Therefore, the curve has to be parametrized, so that it can be written in functional form, $(q, v)^T(t)$, and the parametrization needs

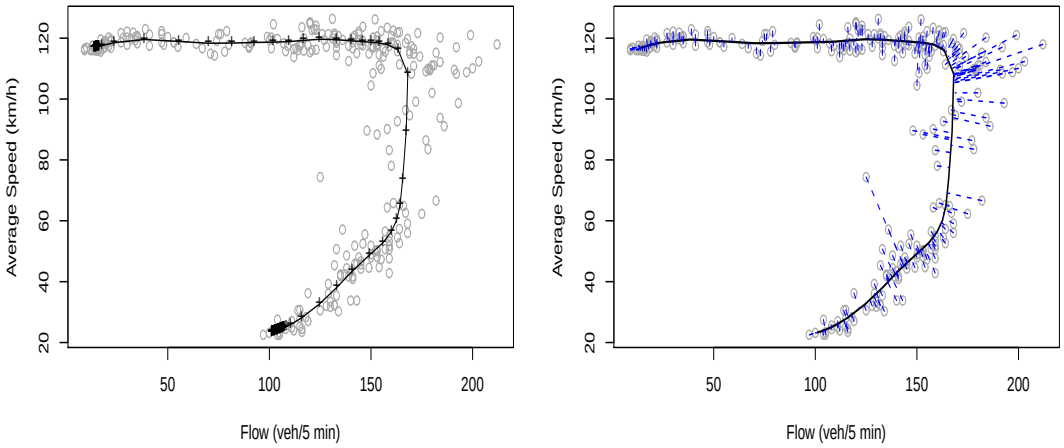


Figure 2: Speed-flow data from 10/07/2007 on Freeway I280-N, VDS 716450; left: with local centers of mass μ^x (+) and local principal curve (solid); right: with orthogonal projections onto the curve (dashed).

to be linked to measurable quantities.

In principle, any parametrization which keeps the right order of data points can be chosen. For example, the most simple one would be an enumeration $t = 1, 2, 3, \ldots$ of all points μ^x starting e.g. from the left top. Another parametrization, which is commonly used for HS principal curves, is the unit-speed parametrization, meaning that distances in the parameter space correspond to distances in the data space. For LPCs, this parametrization would be calculated point by point by adding the Euclidean distance between the current and the last local center of mass, μ^x. An advantage of this parametrization is that it entails an automatic stopping rule in step 5: if the difference between two subsequent parameters falls below a certain threshold the algorithm stops. Parameters between adjacent μ^x are then retrospectively calculated through the arc length of a cubic spline function laid through them (Einbeck, Evers & Hinchliff, 2009).

The benefit of having estimated a function $(q, v)^T(t)$ seems still questionable as we have not measured t. However, note that, from the fundamental identity of traffic flow, we can estimate the average traffic density *along the principal curve* through $k(t) = q(t)/v(t)$, where $(q, v)^T(t)$ are the estimated local centers of mass. Plotting $k(t)$ versus t for the LPC which was fitted in figure 2 (left) leads to figure 3 (left). Here, we see how the traffic density falls monotonically with t (at this occasion $t = 0$ happened to be at the congested end of the principal curve; if the value $t = 0$ was associated to the left top corner then the function $k(t)$ would be monotonically increasing). We call the resulting curve $(t, k(t))$ a *calibration curve*. It can be used to determine the parameter value t from the density k. Then one can use the principal curve to predict speed (v) and flow (q) *simultaneously* from t, and therefore, from k. For instance, for a density $k = 40$ veh/km, one gets from the calibration curve in figure 3 (left) that $\hat{t} = 41.42$, and from the fitted local principal curve the estimates $(\hat{q}, \hat{v})^T(\hat{t}) = (137.94, 41.50)^T$, measured in veh/5min and km/h, respectively.

For almost all speed-flow diagrams which can be represented by a single-branched principal curve, the calibration curve will be monotonic irrespective of the parametrization used. Minor perturbations from this monotonicity will occur if (and only if) the speed-flow data cloud is so strongly skewed that there exists a line through the origin cutting the principal curve twice. This has a purely mathematical reason: points on a line through the origin have a constant ratio $q(t)/v(t)$. Hence, if there is such a line which is crossed twice, then $k(t) = q(t)/v(t)$ will also take this value twice (at non-neighboring values of t), and the monotonicity is broken. This is quite unlikely to happen except at the area around the tails of the principal curve (an example for such a situation is provided in figure 6, bottom right). If necessary, such minor deviations from monotonicity could be smoothed out using a nonparametric smoother with

monotonicity constraint (e.g. Tutz & Leitenstorfer, 2007).

It should be noted that the equilibrium fundamental identity of traffic flow, $k = q/v$, holds only approximately along the principal curve. The reason for this is that its derivation uses a couple of assumptions, one of which being the existence of a number of substreams, in each of which all vehicles are travelling at the same speed (Wardrop, 1952), which is not met as soon as congestion comes in (Hall, 2002). However, as the principal curve is already approximative in nature, this is not a real concern.

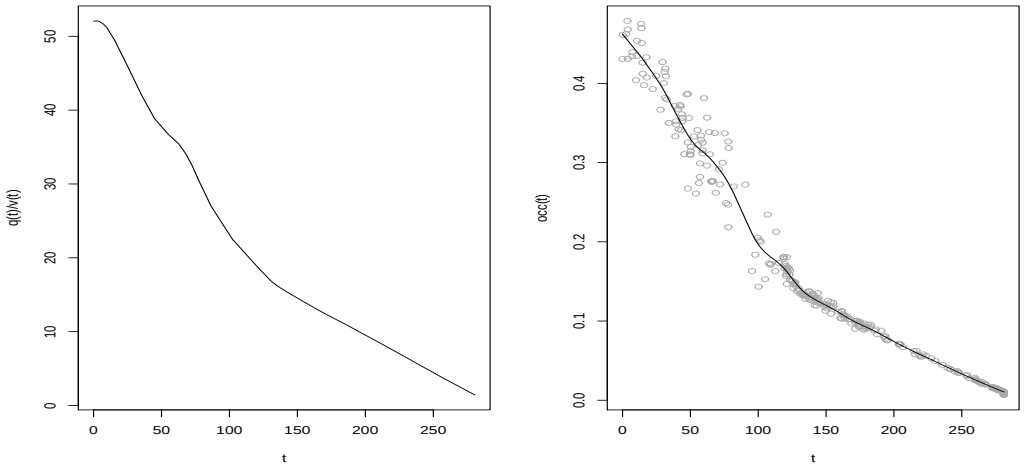


Figure 3: Calibration curves for speed-flow diagram from figure 1; left: using the fundamental identity; right: using measured occupancies.

An alternative to this procedure, which does not require the use of the fundamental identity, is to use the true, observed density, or any other measured quantity which is related to it, for calibration. For instance, we may wish to use the occupancies provided by PeMS for this purpose. We cannot do this directly as the local centers of mass generally do *not* correspond to data points; which implies that we do not have a

measured occupancy for them. However, we do have the measured occupancy for all the data points, and one can compute the projection index t_i of any data point $(q_i, v_i)^T$ projected orthogonally onto the curve. For the data from 10/07/2007, this process is visualized in figure 2 (right). Next, the occupancies are plotted against the projection indices. The resulting data cloud, depicted in figure 3 (right), still has to be smoothed to obtain the calibration curve. We have used penalized splines (Ramsey, 2007) but any other nonparametric smoother could be used. As expected, the smoothed curve is very similar in shape to that one obtained through the fundamental identity. Having now two calibration curves for the same data, one could also use them to convert density to or from occupancy, but we do not pursue this idea further.

3 Analysis of Californian speed-flow diagrams

3.1 Data

Our data has been retrieved from the PeMS 7.3 database (Varaiya, 2004) which records daily traffic data on all major Californian freeways. All reported patterns comprise the time interval from 0.00 to 23.59 of the date stated. The data is collected using loop detectors, i.e. buried coils of wire, whose induction is altered when driven over by a vehicle. Their method is to measure the flow, the number of vehicles that go over a "loop" per unit time, and occupancy, the amount of time each vehicle takes to drive over a loop, of traffic every 30 second period. This is the *raw* data. This is then processed and the average speed, flow and occupancy is calculated for each loop over each 5 minute period. So when looking at a speed-flow diagram, each

point plotted represents the average speed of all cars that passed the loop over 5 minutes and how many of them there were. Collecting the data for each 30 second period separately allows PeMS to detect faulty loops very quickly, thus minimizing the number of "holes", a time period with no data, for which they have to calculate an average using nearby loops and past data. The speed cannot be calculated directly, as only single detector loops are used. Instead, PeMS calculates the "g-factor", a quantity involving the average length of the vehicles on each freeway, in each lane at certain times of the day and week, which combined with the time it takes each one to fully drive over a loop, gives an algorithm with which the speed can be calculated. Details of the PeMS algorithms are provided in Jia, Chen, Coifman & Varaiya (2001). For an account of more recent developments in traffic data collection, see Atluri, Chowdury, Kanhere, Fries, Sarasua & Ogle (2009).

3.2 A temporal sequence of fundamental diagrams

We investigate speed-flow data recorded on the Californian Freeway 1280-N, Lane 1, VDS 716450. The layout of this road around the detector position is shown in figure 4. We consider six consecutive weekdays from Monday 9th of July to Saturday 14th of July 2007 (figure 5), and also six consecutive Tuesdays (figure 6) starting from 26th of June 2007. We have fitted each the local principal curve, and, for the sake of comparison, a HS principal curve and a Greenshields-type curve for each diagram. For the LPC, we use throughout the default parameter values for h, t_0 and x_0 as given in Section 2.1, except for Saturday 14/07/2007 where we used $x_0 = (150, 130)^T$. For the fitting of HS curves, the flow values had to be scaled by the factor 2/3 in order to obtain reasonable results. We observe that, as expected given the results by Chen,

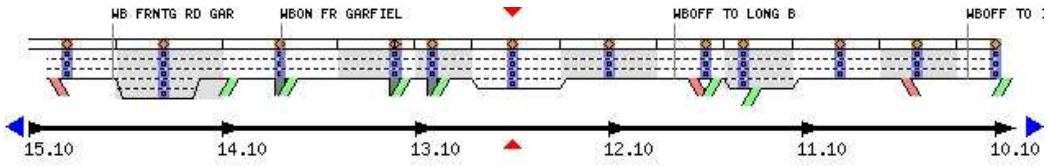


Figure 4: Road layout for freeway 1280-N around VDS 716450 (symbolized through a triangle). Lane 1 is the fastest lane, i.e. the lane in the top. Traffic flows from left to right. Picture taken from PeMS.

Zhang, Tang & Wang (2004), LPCs and HS curves stay generally closer to the data than GM curves. It is also unsurprising that the GM curve produces a useless result for the Saturday, where no congestion was observed. It feels difficult to decide which of HS or LPC gives the better fit. We will argue in the next subsection that the difference between the two curves is rather qualitative than just a quantitative matter of goodness-of-fit.

Modelling speed-flow data through principal curves would only make really sense if the fitted curve is repeatable; i.e. if two principal curves fitted at the same location at different days are similar (under otherwise similar conditions). Therefore, we plot the fitted LPCs for the six weekdays and Tuesdays, respectively, in each one plot and look at their variability. The result is provided in figure 7. We make the following observations: (i) The variability of speed-flow data recorded on consecutive days is larger than the variability between specific weekdays (here Tuesday) over several weeks; (ii) The principal curve representation makes it easy to identify "outlying fundamental diagrams". For instance, in figure 7 (left) the line at the top corresponds clearly to the free-flow situation on Saturday, while in figure 7 (right) the curve drifting away in the bottom-right region is the last of the six Tuesdays, the 31st of July, where the summer holiday season approaches its peak, leading to less congestion during workdays; (iii)

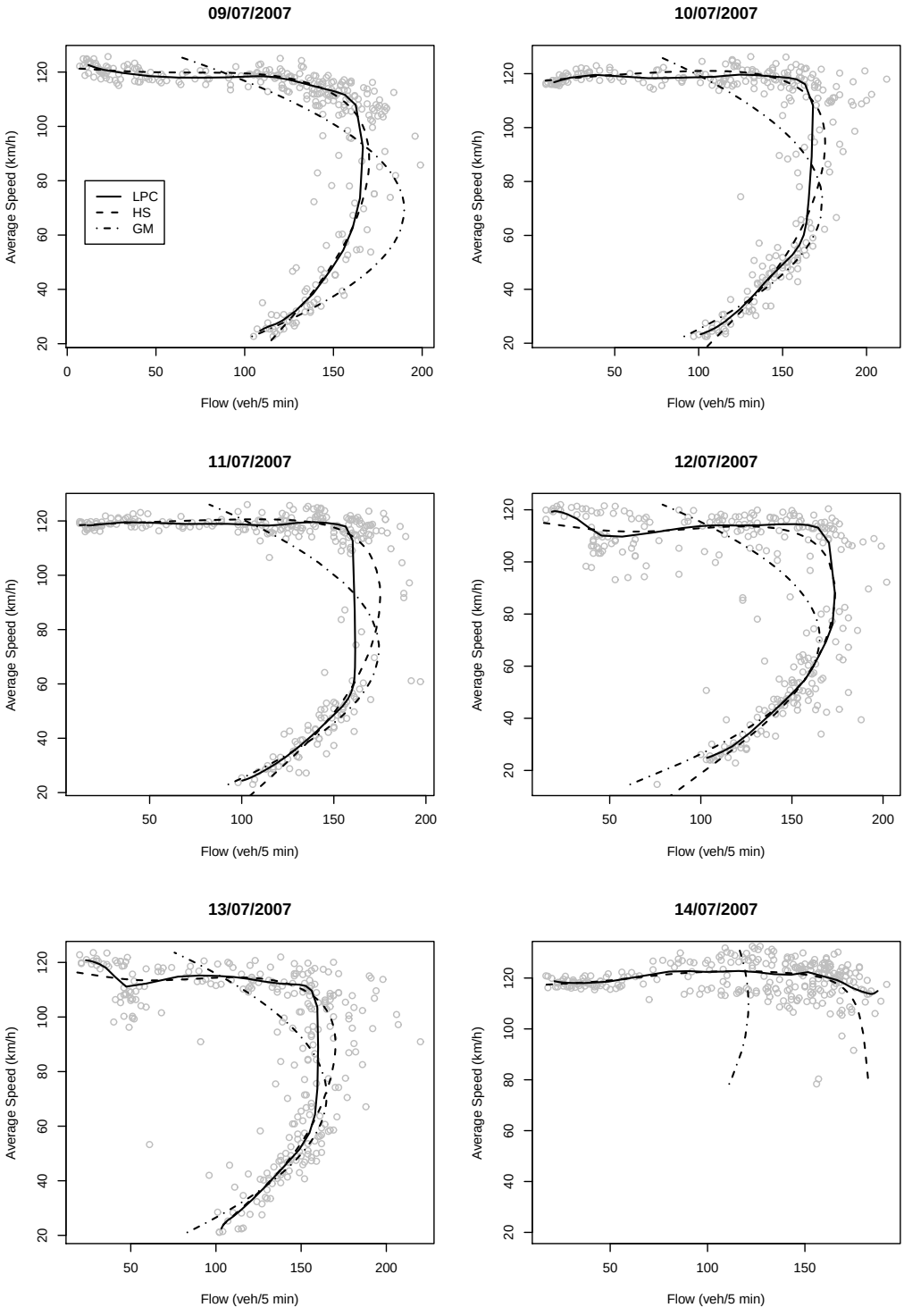


Figure 5: Speed-flow data from VDS 716450 with local principal curve, HS principal curve, and Greenshields curve, for six consecutive weekdays starting 9th of July 2007.

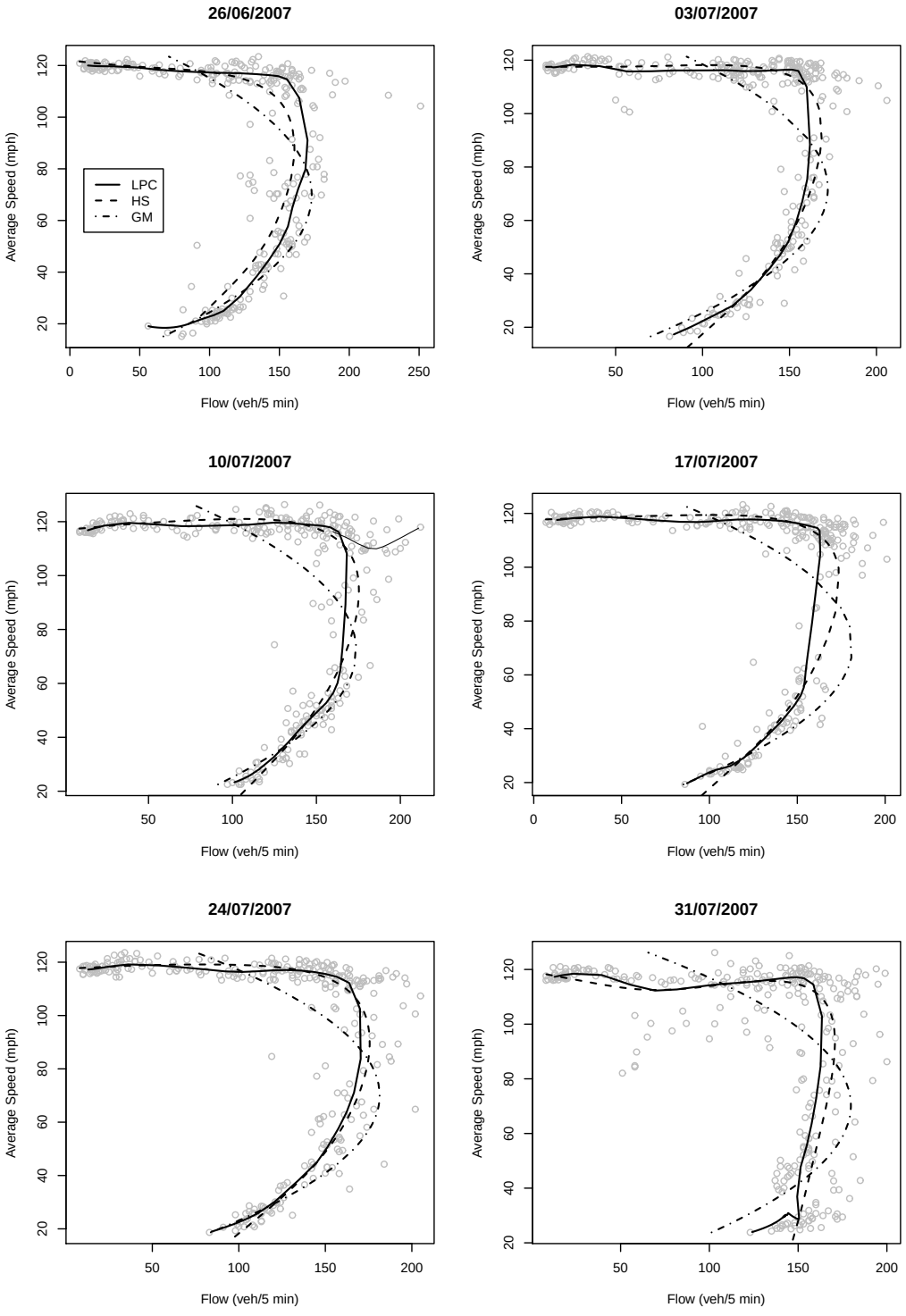


Figure 6: Speed-flow data from VDS 716450 with local principal curve, HS principal curve, and Greenshields curve, for six consecutive Tuesdays starting 26th of June 2007.

Otherwise, the shape and position of the principal curves does not vary substantially. Hence, if one has a "reference principal curve" for a certain road then this may be used to predict the traffic behaviour in the future.

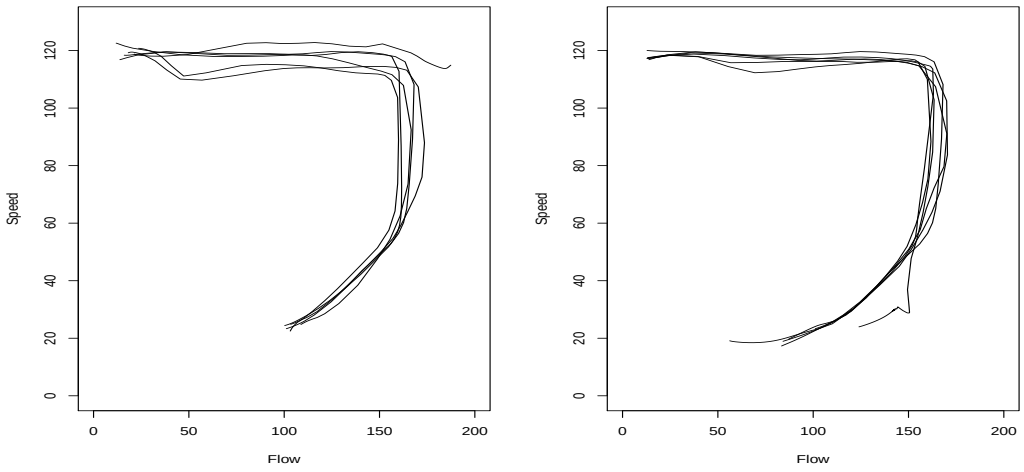


Figure 7: Speed-flow data from VDS 716450 with fitted LPCs; left: six consecutive weekdays; right: six consecutive Tuesdays.

3.3 Capacity, and capacity drop

The capacity of a road at a particular location is usually defined as the maximum flow rate q_{max} achievable under equilibrium conditions (Wu, 2002). The equilibrium road capacity can be read from the fundamental diagram by taking that point of the fitted curve which corresponds to a maximum flow rate. Table 1 lists the road capacities obtained for the six consecutive Tuesdays considered earlier, for all three methods of estimation. For the LPC and HS technique, we use the maximum over all flow rates given in the local center of masses, or knots of the fitted curve, respectively. For the

Table 1: Estimated capacities in veh/hour for Lane 1 of freeway I-280N on six consecutive Tuesdays, with means and standard deviations of the six estimates given underneath.

q_{max}	LPC	HS	GM
26/06/2007	2041.34	1929.08	2077.41
03/07/2007	1941.19	2021.51	2064.37
10/07/2007	2017.30	2103.89	2084.48
17/07/2007	1956.83	2082.53	2177.26
24/07/2007	2045.29	2107.66	2176.81
31/07/2007	1961.09	2048.67	2156.70
$\overline{q_{max}}$	1993.84	2048.89	2122.84
$s(q_{max})$	46.19	67.40	52.87

Greenshields model, we take the maximum of the fitted parabola $q = bv + cv^2$, which is attained at $v = -b/2c$. For a better comparability with other literature we provide the results in veh/hour and not veh/5 min, i.e. we multiply all obtained values by 12. We observe that the Greenshields model produces the highest capacities; they are overestimated as also observed by Wu (2002). Both HS and LPC produce road capacities (or better, in our context: lane capacities) far below the GM values. It is not clear whether the HS or LPC capacities are closer to the true capacities, but the LPC estimates are at least more precise and repeatable as their standard deviation is considerably smaller. Interestingly, they are even less variable than the capacities obtained from GM – one would expect a parametric model to be superior at least in terms of variability!

Of particular interest for traffic engineers is the capacity at bottlenecks, i.e. locations where the total arriving flow reaches or exceeds the downstream capacity, for instance due to a merging of lanes (Papageorgiou, Papamichail, Spiliopoulou & Lentzakis, 2008). This creates congestion, and the exiting flow is reduced below the downstream capacity of the bottleneck. This phenomenon is usually referred to as capacity drop. We see from Table 1 that the estimated capacity for LPC was generally smaller than that estimated for HS or GM. The sole exception for this is the data from 26/06/2007; however here the HS fit was clearly suboptimal and it seems as if it was not fully converged. From figure 6 we observe that the fitted LPC falls much sharper compared to their HS or GM counterpart, a behaviour which is in line with literature stating that there should be a vertical drop of speed (Hall, Hurdle & Banks, 1992). This rapid decrease happens where the limit to the number of vehicles that can be driving on the freeway without affecting their average speed is reached. A possible conclusion is that the LPC accounts more accurately for a capacity drop, i.e. its maximal flow value estimates directly the reduced capacity due to capacity drop rather than the maximum output flow.

Taking it strictly, speed-flow data with capacity drop are not only curved but also branched, as at some occasions vehicles manage to maintain their speed at high flow values during queue formation. The point at which the curve starts to bend can be close to the highest flow value as e.g. on 09/07/2007, or at any point along the freeflow section of the data cloud, see also figure 8 (a). The existence of such a sidewards branch has already been reported by Hall, Hurdle & Banks (1992) in their schematic speed-flow diagram in figure 2. Unlike HS curves, local principal curves provide the methodology to deal with this. The simple solution is to determine two starting points by hand, one in the congested and one in the uncongested part, fit each

branch as usual, and overlay the curves. This is exemplarily illustrated through a thin curve added to the plot for the data from 10/07/2007 in figure 6. One can think of the flow range covered by the thin curve as the extent of the capacity drop. We note finally that for local principal curves an automatic procedure to detect the positions of the bifurcations is available. A full description of this technique would burst the framework of this paper, for details see Einbeck, Tutz & Evers (2005a).

3.4 Selected representative patterns

Here we provide a list of further interesting patterns which we encountered during our studies. Some are just special variations of the fundamental diagram, while others have an entirely different shape. We do not claim our list to be exhaustive but it is representative of the full range of patterns we found in an extensive search. We have fitted a local principal curve to each of them, using at several occasions multiple starting points in order to cover all branches. Due to the particular shape of the patterns, most of the starting points are different from our default value given in Section 2.1, but we will in general just report in which branch they were located as this is all that is relevant. We use again $h = t_0 = 12$ unless stated otherwise, but $h = t_0$ in any case. We do not compare our results to HS or GM as these methods are not able to deal with branched patterns.

Two regimes. The data shown in figure 8 (a) was collected on July 11th 2007 on Lane 1 of Freeway SR57-N. This is, in principle, nothing else than a usual fundamental diagram, just that the time between reaching the flow limit of the freeway and severe congestion is much smaller, whereas previously this change happened gradually, with

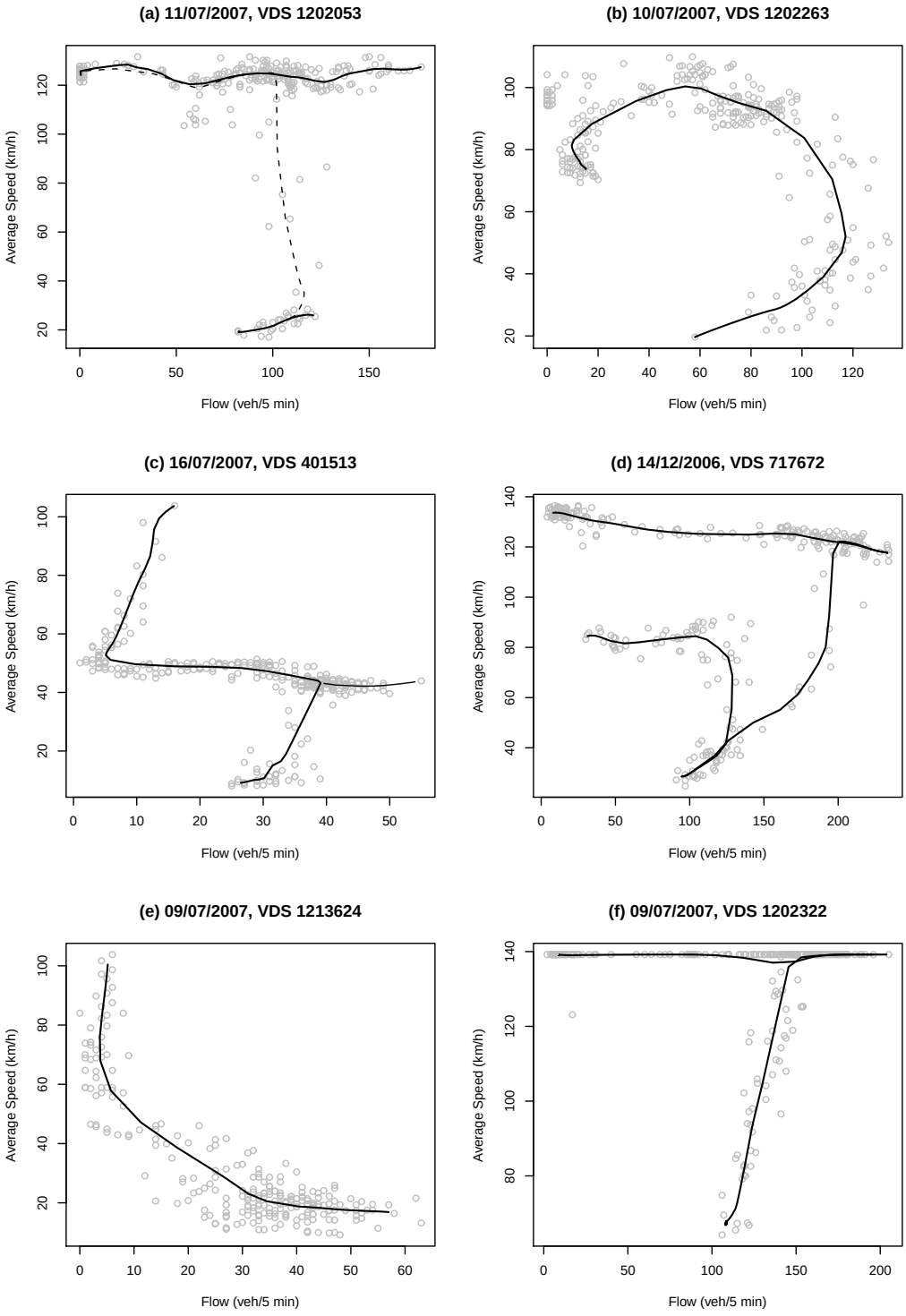


Figure 8: Representative speed-flow patterns collected from PeMS, with fitted local principal curves. The corresponding road layouts are provided in figure 9.

the flow rate and speed decreasing over perhaps one or two hours. This situation could occur on a freeway frequented by people travelling to or from work, or due to an accident just after the loop, as the number of vehicles on the freeway would then increase almost instantaneously at a certain time. Taking the viewpoint that these are two disjoint regimes (without a genuine transition region) which cannot be described by a single-branched curve, one could fit separate branches using appropriate starting points in each cluster (see solid lines in figure 8 (a)). This requires to use a relatively small bandwidth, here $h = 6$. If one takes the view that this is a usual fundamental diagram, one can increase the bandwidth to our usual setting $h = 12$, which gives the dashed line, and the typical τ-shaped curve is obtained, which is now similar to that one in figure 6 on 10/07/2007, just featuring a more radical capacity drop.

Night time drivers. When considering any speed-flow relationship, it is often observed that the freeflow data with a small flow rate decreases in speed also as shown in figure 8 (b), whose data was collected on Freeway SR57-N, Lane 5, on July 10th 2007. An explanation for this is that free flow at a very low flow rate is likely to occur very late at night when the freeways are at their least busy, therefore many road users will choose to drive at a slightly slower speed than they might do at such a flow rate during the day. This effect is especially visible on very busy freeways as then the only very low flow rates during freeflow would be late at night, however on a quieter freeway very low flow rates are much more likely to occur during the day too, so instead of a curve like that shown above, the left end of the freeflow section might get thicker, representing very low flow rates both in the day, at a high speed, and the night, at a lower speed. This data is extremely curved, however the LPC, using a randomly chosen starting point, does model it very closely.

Undertaking. Another different effect that can be observed in the slowest lane of a freeway is that of undertaking. The typical speed-flow curve with a slow free flow speed can be seen in the data cloud given in figure 8 (c) from Freeway I80-W, Lane 5, on July 16th 2007. Additionally, the curve has average speeds much higher than the freeflow speed at low flow rates. This can be explained by considering that slower lanes may be used for undertaking other vehicles, in which case a much higher speed than the average free flow speed of the heavy vehicles which generally occupy this lane must be used. It is impractical to attempt to undertake when this lane is busy, when the flow is high, which is why the effect is isolated to a small range of flow rate only. This data cloud is modelled using two starting points. Specifically, one starting point in the bottom cluster launched a curve which covered almost the entire data cloud (thick line). A second curve with starting point $x_0 = (35, 50)^T$ was fitted for the remaining data points associated to very high flow values (thin line).

Change in driving conditions. Figure 8 (d) represents the type of data cloud that occurs when there has been some change in the driving conditions and was taken on Freeway I210-E, Lane 1, on Dec 14th 2006. Adverse weather or road works are two examples of incidents that can cause this kind of change. In such circumstances, road users will generally decrease their choice of freeflow speed in order to have better control of their vehicle. During a period of congestion, however, the decrease in average speed is much smaller as it is very slow anyway. Therefore, if the data is collected over a period of time in which we have both normal road conditions and adverse weather conditions and both freeflow and congestion has occurred within both, the data cloud will have the appearance shown in the graph, with two branches (or regimes), representing the freeflow at the two different average speeds, connected to the congestion curves which are approximately the same under both conditions. This

effect is discussed in detail in Huang & Ran (2003). The LPC is fitted using three starting points, i.e. one on each branch, and a slightly lower bandwidth than usual, $h = 11$.

Slow lanes. Data taken from the slowest lanes on a freeway can take a different shape than the faster lanes. The data shown in figure 8 (e), which was taken from Freeway SR57-S, Lane 5, on July 9th 2007, has its smallest flow rate at the highest average speed and then increases in flow inversely to the speed. This effect could be explained by considering the fact that heavy vehicles have a bigger stopping distance, so have to leave a bigger distance between themselves and the vehicle in front of them. Therefore it is possible that the drivers slow down at a much lower flow rate compared to the previous situations. The reduced speed implies a reduced stopping distance, allowing each vehicle to become closer to the one in front, which in turn leads to increasing flow. Of course, if the slow lane became more congested the flow would then decrease with the speed as for a usual speed-flow pattern. The LPC is easily fitted for this data using any randomly chosen starting point. Since this curve is unbranched and cut only once by any straight line through the origin, it could also be calibrated.

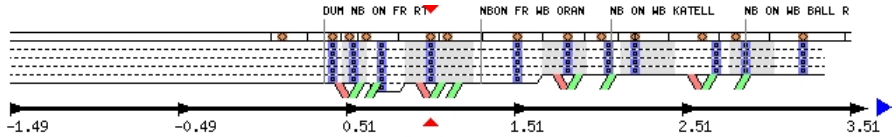
Faulty data. The data shown in figure 8 (f), taken on Freeway SR57-S, Lane 1, VDS number 1202322 during July 9th 2007, is an example demonstrating the occurrence of unusual data possibly caused by one or more faulty VDS detector as discussed in Pescovitz (2001). In general these faults in the PeMS database are identified and "corrected", either by looking at the corresponding data from a series of nearby working VDS detectors, looking at the history of that VDS detector, or combining the two. There are some cases, of course, where faulty data is overlooked. A common type of

faulty data is a long period of time in which exactly the same average speed is recorded for every five minute period. The local principal curve has fitted this data by using one starting point in each of the two branches.

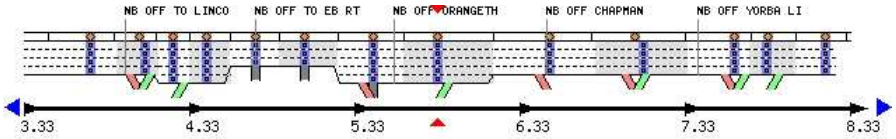
4 Conclusion

We have investigated local principal curves as a tool for modelling of speed-flow relationships. We argued that neither a functional relationship of type $v = v(q)$ nor $q = q(v)$ accounts for the true nature of the data, which is that actually both variables can be considered as being generated simultaneously from some underlying distribution, and that they are jointly driven by a third variable: the traffic density.

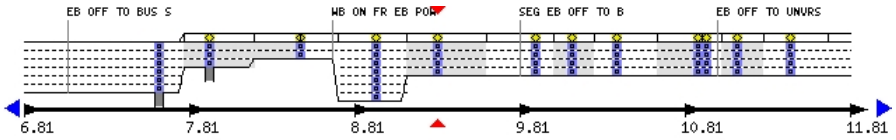
Principal curves are an attractive tool for the modelling of speed-flow data which take this symmetric association between both variables into account. They treat both variables as jointly generated depending on some underlying parameter t, which we have shown to be closely and monotonically related to the traffic density. As such, we could consider the parametrization in itself as another measure of "traffic concentration", a broader term suggested by Hall (2002) encompassing both density and occupancy. The relationship between parameter and density can be quantified through a calibration curve, an approximate version of which can be generated even without knowledge of the traffic density via the fundamental identity of traffic flow (in principle, also an external "reference" calibration curve, which would form a characteristic of the road under certain default conditions, could be used instead, if one standardizes the "0" value of the parametrization).



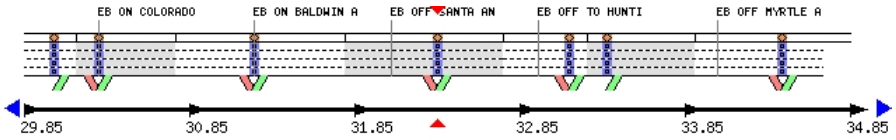
(a)



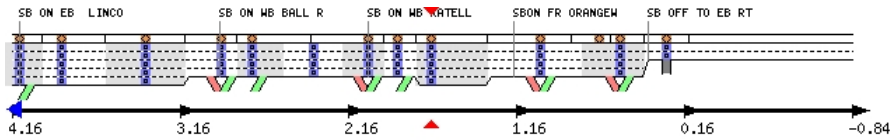
(b)



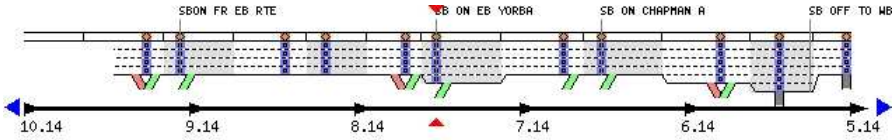
(c)



(d)



(e)



(f)

Figure 9: Road layouts corresponding to each pattern considered in Section 3.4. Traffic flows from left to right. Lane 1 is the fastest ("top") lane in each panel. The detectors are positioned in the center of each extract. The pictures are taken from PeMS.

For the estimation of the principal curve, a variety of algorithms are available — however, the implementation should provide access to the parametrization, and the curve should be smooth rather than a polygonal line, as the projected points will cluster around the edges of the segments otherwise. This leaves us essentially with the two methods applied in this paper, the HS curves and the LPCs, or variants of them. We do not want to use this paper to make the case that either of them gives a better goodness of fit than the other one. Other papers have carried out investigations of this type (Delicado & Huerta, 2003, Einbeck, Tutz & Evers, 2005b), but the results should not be over-valued as in all principal curve algorithms there are quite a few tuning parameters and options to choose, and any attempt to decide for a best principal curve algorithm based on such an arbitrary choice seems elusive. However, we have tried to argue that LPCs and HS curves indeed show an intrinsically different behavior at flow values close to the road capacity, which is not just a matter of goodness-of-fit. In Section 3.4 we gave some examples of more complex patterns (branched, disconnected), which, by definition, could not be dealt with using HS curves. One should also note that it has been frequently demonstrated (e.g. Delicado & Huerta, 2003) that HS principal curves can heavily fail even for single-branched connected clouds when the data structure is strongly bent or twisted (e.g. a spiral). The problem is that they are based on an initially estimated line (usually the first linear principal component line). If the order of projection indices of the data projected onto this line is very different from the projection indices of the data onto the "ideal" principal curve (i.e. a smooth curve which would capture the structure of the entire data cloud), then this cannot be corrected in later iterations of the algorithm. Advantages of LPCs are their simplicity (a simple alternation between calculation of the local mean and the localized first principal component), the availability of additional tools for projection and prediction (which does not mean that this couldn't be implemented for HS curves

as well), and their flexibility to adapt to very complex data structures. This comes at the price of somewhat larger variability compared to other principal curve algorithms, and the choice of the starting point x_0 can play quite a crucial role for some data sets. To make sure that the reader is able to reproduce all results presented in this work, we have put the R source code of all data analyses carried out in this paper at

`http://www.maths.dur.ac.uk/~dma0je/lpc/lpc.htm`,

along with the R function `lpc` itself.

We have also provided an overview of types of speed-flow diagrams, based on an extensive search of patterns collected from the PeMS database, which may be useful for further reference. Our intention was not to treat all mathematical and technical details in full depth. The nature and value of calibration curves for multi-regime situations as in Section 3.4 has not been investigated yet and deserves further attention. We finish with a word of caution: Though principal curves may be suitable for descriptive and comparative purposes, and to a certain extent also for prediction, they cannot account for the complex traffic dynamics and stochastic processes underlying particularly the congestion and transition regions. For instance, even if one manages to calibrate a curve as in figure 8 (d), it does not seem very likely that this provides a reliable reference for the behavior of the road in the future. A thorough investigation of congested traffic states taking the complex spatiotemporal behavior of traffic data into account has been provided by Schönhof & Helbing (2007). More recently, Sheu, Lan & Huang (2009) discussed a novel real-time recurrent learning algorithm for short-term prediction of traffic dynamics.

Acknowledgments

The second author of this paper was supported by Nuffield Research Bursary URB/34397. We are grateful to M.Z.F. Li, Nanyang Business School, Singapore, for helpful comments at the draft stage of this paper, and M. Zayed, Durham University, UK, for his contribution to the R implementation of the `lpc` function.

References

Atluri, M., Chowdury, M., Kanhere, N., Fries, R., Sarasua, W., and Ogle, J. (2009). Development of a sensor system for traffic data collection. *Journal of Advanced Transportation* **43**, 1–20.

Chen, D., Zhang, J., Tang, S., and Wang, J. (2004). Freeway traffic stream modeling based on principal curves and its analysis. *IEEE Transactions on Intelligent Transportation Systems* **5**, 246–258.

Delicado, P. and Huerta, M. (2003). Principal curves of oriented points: Theoretical and computational improvements. *Computational Statistics* **18**, 293–315.

Drake, L. S., Schoefer, J. L., and May, A. D. (1967). A statistical analysis of speed-density hypotheses. *Highway Research Record* **154**, 53–87.

Einbeck, J., Evers, L., and Hinchliff, K. (2009). Data compression and regression based on local principal curves. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in Data Analysis, Data Handling and Business Intelligence*, Heidelberg, pp. 701–712. Springer.

Einbeck, J., Tutz, G., and Evers, L. (2005a). Exploring multivariate data structures

with local principal curves. In C. Weihs & W. Gaul (Eds.), *Classification - The Ubiquitous Challenge.*, Heidelberg, pp. 257–263. Springer.

Einbeck, J., Tutz, G., and Evers, L. (2005b). Local principal curves. *Statistics and Computing* **15**, 301–313.

Greenshields, B. D. (1935). A study of traffic capacity. *Highway Research Board Proc.* **14**, 448–477.

Hall, F. L. (2002). Traffic stream characteristics. In N. H. Gartner, C. J. I. Messer, & A. Rathi (Eds.), *Traffic flow theory: A state-of-the-art report.* Georgetown Pike, VA: Turner Fairbank Highway Research Center (TFHRC).

Hall, F. L., Hurdle, V. F., and Banks, J. M. (1992). Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relations on freeways. *Transportation Research Record* **1365**, 12–17.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84**, 502–516.

Huang, S. H. and Ran, B. (2003). An application of neural network on traffic speed prediction under adverse weather condition. In *TRB 2003 Annual Meeting CD-ROM. www.ltrc.lsu.edu/TRB_82/TRB2003-000915.pdf.*

Jia, Z., Chen, C., Coifman, B., and Varaiya, P. (2001). The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors. *Proceedings of the IEEE 4th International ITS Conference.*

Kockelman, K. M. (2001). Modeling traffic's flow-density relation: Accomodation of multiple flow regimes and traveler types. *Transportation* **24**, 363–374.

Li, M. Z. F. (2008). A generic characterization of equilibrium speed-flow curves. *Transportation Science* **42**, 220–235.

Papageorgiou, M., Papamichail, I., Spiliopoulou, A., and Lentzakis, A. (2008). Real-time merging traffic control with applications to toll plaza and work zone man-

agement. *Transportation Research Part C* **16**, 535–553.

Pescovitz, D. (2001). The mathematics of high-tech highways. *ScienceMatters@Berkeley* **1**, Issue 2.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramsey, J. (2007). *pspline: Penalized Smoothing Splines*. R package version 1.0-12. R port by Brian Ripley.

Schönhof, M. and Helbing, D. (2007). Empirical features of congested traffic states and their implications for traffic modelling. *Transportation Science* **41**, 135–166.

Sheu, J.-B., Lan, L., and Huang, Y.-S. (2009). Short-term prediction of traffic dynamics with real-time recurrent learning algorithms. *Transportmetrica* **5**, 59–83.

Tutz, G. and Leitenstorfer, F. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8**, 654–673.

Varaiya, P. (2004). Freeway performance measurement system (PeMS) version 4. Technical Report 31, California Partners for Advanced Transit and Highways (PATH).

Wardrop, J. G. (1952). Some theorectical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II* **1**, 325–362.

Wu, N. (2002). A new approach for modelling of fundamental diagrams. *Transportation Research Part A* **36**, 867–884.

Xia, J. and Chen, M. (2007). A nested clustering technique for freeway operating condition classification. *Computer-Aided Civil and Infrastructure Engineering* **22**, 430–437.